An Annotated Bibliography of Privacy-Violating Attacks and Related Work

The Research And Methodology Directorate

Version 1.0

By Philip Leclerc and Pavel Zhuravlev

Issued January 23, 2023

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

An Annotated Bibliography of Privacy-Violating Attacks and Related Work*

Philip Leclerc, Pavel Zhuravlev (USCB Research and Methodology Directorate)

Introduction

This annotated bibliography is intended to serve as a convenient guide to papers constituting the scientific literature on privacy-violating attacks. Attacks with a modern "reconstruction-reidentification" flavor were the original motivation for collating this bibliography, but its scope has since been expanded to include references to more general kinds of privacy-violating attacks, to papers describing the theoretical foundations of possible attacks, to papers focused on teaching and establishment of standards related to privacy-violating attacks, to papers on the design and implementation of the 2020 Census Disclosure Avoidance System (DAS), the construction of which was motivated in significant part by awareness of the literature for which this bibliography serves as a guide, and, lastly, to papers focused on inference about a target person's membership in a data set. Papers are divided into five categories, following these descriptions; divisions between categories are not exact, but are meant to roughly capture the overall spirit of each paper.

Note that, in the categories detailing attacks, we also include some papers with primarily negative results (e.g., showing or arguing that certain attack vectors are ineffective against certain classes of database or disclosure protections), though this work is in the minority in each list. However, we do not include papers that are strictly about a particular kind of defense against attacks (e.g., we do not include papers that study k-anonymous methods, nor papers exclusive to the study of differentially private methods, except in the case of category 4, where we include the small number of papers required to understand the design of the 2020 DAS). In this sense, this bibliography focuses on the kinds of attacks that have been designed, not on the kinds of disclosure-avoidance defenses that researchers have suggested.

This bibliography is unlikely to ever be truly exhaustive, and is a long-term work-in-progress, but we developed it with thorough coverage of the literature as our primary goal.

1 Papers Demonstrating Empirical Attacks

Our first category of papers is the most straightforward: these include at least one example of an empirical demonstration of a privacy-violating attack. Note that many of these papers also describe the theoretical foundations of their attacks.

*This paper represents the views of the authors not the U.S. Census Bureau.

Empirical Attacks

[1] Kerem Ayoz, Erman Ayday, and Ercument Cicek. "Genome Reconstruction Attacks Against Genomic Data-Sharing Beacons". In: *Proceedings on Privacy Enhancing Technologies*. 2021. DOI: 10.2478/popets-2021-0036. URL: https://pubmed.ncbi.nlm.nih.gov/34746296/.
Sharing genome data in a privacy-preserving way stands as a major bottleneck in front of the scientific progress promised by the big data era in genomics. A community-driven protocol named genomic data-sharing beacon protocol has been widely adopted for sharing genomic data. The system aims to provide a secure, easy to implement, and standardized interface for data sharing by only allowing yes/no queries on the presence of specific alleles in the dataset. However, beacon protocol was recently shown to be vulnerable against membership inference attacks. In this paper, we show that privacy threats against genomic data sharing beacons are not limited to membership inference. We identify and analyze a novel vulnerability of genomic data-sharing beacons: genome reconstruction. We show that it is possible to successfully reconstruct a substantial part of the genome of a victim when the attacker knows the victim has been added to the beacon in a recent update. In particular, we show how an attacker can use the inherent correlations in the genome and clustering techniques to run such an attack in an efficient and accurate way. We also show that even if multiple individuals are added to the beacon during the same update, it is possible to identify the victim's genome with high confidence using traits that are easily accessible by the attacker (e.g., eye color or hair type). Moreover, we show how a reconstructed genome using a beacon that is not associated with a sensitive phenotype can be used for membership inference attacks to beacons with sensitive phenotypes (e.g., HIV+). The outcome of this work will guide beacon operators on when and how to update the content of the beacon and help them (along with the beacon participants) make informed decisions.

[2] Erik Buchholz et al. "Reconstruction Attack on Differential Private Trajectory Protection Mechanisms". In: *Annual Computer Security Applications Conference*. 2022, pp. 279–292. URL: https://dl.acm.org/doi/abs/10.1145/3564625.3564628.
Location trajectories collected by smartphones and other devices represent a valuable data source for applications such as location-based services. Likewise, trajectories have the potential to reveal sensitive information about individuals, e.g., religious beliefs or sexual orientations. Accordingly, trajectory datasets require appropriate sanitization. Due to their strong theoretical privacy guarantees, differential private publication mechanisms receive much attention. However, the large amount of noise required to achieve differential privacy yields structural differences, e.g., ship trajectories passing over land. We propose a deep learning-based Reconstruction Attack on Protected Trajectories (RAoPT), that leverages the mentioned differences to partly reconstruct the original trajectory from a differential private release. The evaluation shows that our RAoPT model can reduce the Euclidean and Hausdorff distances between the released and original trajectories by over 68% on two real-world datasets under protection with $\varepsilon \leq 1$. In this setting, the attack increases the average Jaccard index of the trajectories' convex hulls, representing a user's activity space, by over 180%. Trained on the GeoLife dataset, the model still reduces the Euclidean and Hausdorff distances by over 60% for T-Drive trajectories protected with a state-of-the-art mechanism ($\varepsilon = 0.1$). This work highlights shortcomings of current trajectory publication mechanisms, and thus motivates further research on privacy-preserving publication schemes.

[3] Joseph Calandrino et al. ""You Might Also Like:" Privacy Risks of Collaborative Filtering". In: *Proceedings of the 2011 IEEE Symposium on Security and Privacy*. 2011, pp. 231–246. DOI: DOI: 10.1109/SP.2011.40. URL: https://www.cs.utexas.edu/~shmat/shmat_oak11ymal.pdf.
Many commercial websites use recommender systems to help customers locate products and content. Modern recommenders are based on collaborative filtering: they use patterns learned from users' behavior to make recommendations, usually in the form of related-items lists. The scale and complexity of these systems, along

with the fact that their outputs reveal only relationships between items (as opposed to information about users), may suggest that they pose no meaningful privacy risk. In this paper, we develop algorithms which take a moderate amount of auxiliary information about a customer and infer this customerś transactions from temporal changes in the public outputs of a recommender system. Our inference attacks are passive and can be carried out by any Internet user. We evaluate their feasibility using public data from popular websites Hunch, Last.fm, LibraryThing, and Amazon.

[4] Aloni Cohen and Kobbi Nissim. "Linear Program Reconstruction in Practice". In: *arXiv preprint* (2018). URL: https://doi.org/10.48550/arXiv.1810.05692.
We briefly report on a successful linear program reconstruction attack performed on a production statistical queries system and using a real dataset. The attack was deployed in test environment in the course of the Aircloak Challenge bug bounty program and is based on the reconstruction algorithm of Dwork, McSherry, and Talwar. We empirically evaluate the effectiveness of the algorithm and a related algorithm by Dinur and Nissim with various dataset sizes, error rates, and numbers of queries in a Gaussian noise setting.

[5] Cynthia Dwork et al. "Robust Traceability from Trace Amounts". In: *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS '15)*. ACM Digital Library, 2015, pp. 650–669. URL: https://doi.org/10.1109/FOCS.2015.46.
The privacy risks inherent in the release of a large number of summary statistics were illustrated by Homer et al. (PLoS Genetics, 2008)[11], who considered the case of 1-way marginals of SNP allele frequencies obtained in a genome-wide association study: Given a large number of minor allele frequencies from a case group of individuals diagnosed with a particular disease, together with the genomic data of a single target individual and statistics from a sizable reference dataset independently drawn from the same population, an attacker can determine with high confidence whether or not the target is in the case group. In this work we describe and analyze a simple attack that succeeds even if the summary statistics are significantly distorted, whether due to measurement error or noise intentionally introduced to protect privacy. Our attack only requires that the vector of distorted summary statistics is close to the vector of true marginals in L1 norm. Moreover, the reference pool required by previous attacks can be replaced by a single sample drawn from the underlying population. The new attack, which is not specific to genomics and which handles Gaussian as well as Bernouilli data, significantly generalizes recent lower bounds on the noise needed to ensure differential privacy (Bun, Ullman, and Vadhan, STOC 2014, Steinke and Ullman, 2015), obviating the need for the attacker to control the exact distribution of the data.

[6] Cynthia Dwork et al. "Exposed! A Survey of Attacks on Private Data". In: *Annual Review of Statistics and Its Application* (2017). DOI: 10.1146/annurev-statistics-060116-054123. URL: https://privacytools.seas.harvard.edu/publications/exposed-survey-attacks-private-data.
Privacy-preserving statistical data analysis addresses the general question of protecting privacy when publicly releasing information about a sensitive dataset. A privacy attack takes seemingly innocuous released information and uses it to discern the private details of individuals, thus demonstrating that such information compromises privacy. For example, re-identification attacks have shown that it is easy to link supposedly de-identified records to the identity of the individual concerned. This survey focuses on attacking aggregate data, such as statistics about how many individuals have a certain disease, genetic trait, or combination thereof. We consider two types of attacks: reconstruction attacks, which approximately determine a sensitive feature of all the individuals covered by the dataset, and tracing attacks, which determine whether or not a target individual's data are included in the dataset. We also discuss techniques from the differential privacy literature for releasing approximate aggregate statistics while provably thwarting any privacy attack.

[7] Francesca Falzon et al. “Full Database Reconstruction in two Dimensions”. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 2020, pp. 443–460. URL: https://doi.org/10.1145/3372297.3417275.
In the past few years, we have seen multiple attacks on one-dimensional databases that support range queries. These attacks achieve full database reconstruction by exploiting access pattern leakage along with known query distribution or search pattern leakage. We are the first to go beyond one dimension, exploring this threat in two dimensions. We unveil an intrinsic limitation of reconstruction attacks by showing that there can be an exponential number of distinct databases that produce equivalent leakage. Next, we present a full database reconstruction attack. Our algorithm runs in polynomial time and returns a poly-size encoding of all databases consistent with the given leakage profile. We implement our algorithm and observe real-world databases that admit a large number of equivalent databases, which aligns with our theoretical results.

[8] Danilo Favato et al. “A Novel Reconstruction Attack on Foreign-trade Official Statistics, With a Brazilian Case Study”. In: *Proceedings on Privacy Enhancing Technologies* (4 2022). DOI: 10.56553/popets-2022-0124. URL: https://petsymposium.org/popets/2022/popets-2022-0124.php.
In this paper we describe, formalize, implement, and experimentally evaluate a novel transaction re-identification attack against official foreign-trade statistics releases in Brazil. The attack’s goal is to re-identify the importers of foreign-trade transactions (by revealing the identity of the company performing that transaction), which consequently violates those importers’ fiscal secrecy (by revealing sensitive information: the value and volume of traded goods). We provide a mathematical formalization of this fiscal secrecy problem using principles from the framework of quantitative information flow (QIF), then carefully identify the main sources of imprecision in the official data releases used as auxiliary information in the attack, and model transaction re-construction as a linear optimization problem solvable through integer linear programming (ILP). We show that this problem is NP-complete, and provide a methodology to identify tractable instances. We exemplify the feasibility of our attack by performing 2,003 transaction re-identifications that in total amount to more than $137M, and affect 348 Brazilian companies. Further, since similar statistics are produced by other statistical agencies, our attack is of broader concern.

[9] Michael Fire et al. “Links Reconstruction Attack”. In: *Security and Privacy in Social Networks*. Ed. by Yaniv Altshuler et al. New York, NY: Springer New York, 2013, pp. 181–196. ISBN: 978-1-4614-4139-7. DOI: 10.1007/978-1-4614-4139-7_9. URL: https://doi.org/10.1007/978-1-4614-4139-7_9.
The explosion in the use of social networks has also created new kinds of security and privacy threats. Many users are unaware of the risks involved with exposing their personal information, which makes social networks a “bonanza” for identity thieves. In addition, it has already been proven that even concealing all personal data might not be sufficient for providing protection, as personal information can be inferred by analyzing a person’s connections to other users. In attempts to cope with these risks, some users hide parts of their social connections to other users. In this paper we present “link reconstruction attack”, a method that can infer a user’s connections to others with high accuracy. This attack can be used to detect connections that a user wanted to hide in order to preserve his privacy. We show that concealing one’s links is ineffective if not done by others in the network. We also provide an analysis of the performances of various machine learning algorithms for link prediction inside small communities.

[10] Sébastien Gambs, Ahmed Gmati, and Michel Hurfin. “Reconstruction Attack Through Classifier Analysis”. In: *IFIP Annual Conference on Data and Applications Security and Privacy*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 274–281. DOI: https://doi.org/10.1007/

978-3-642-31540-4_21. URL: https://link.springer.com/chapter/10.1007/978-3-642-31540-4_21.

In this paper, we introduce a novel inference attack that we coin as the reconstruction attack whose objective is to reconstruct a probabilistic version of the original dataset on which a classifier was learnt from the description of this classifier and possibly some auxiliary information. In a nutshell, the reconstruction attack exploits the structure of the classifier in order to derive a probabilistic version of dataset on which this model has been trained. Moreover, we propose a general framework that can be used to assess the success of a reconstruction attack in terms of a novel distance between the reconstructed and original datasets. In case of multiple releases of classifiers, we also give a strategy that can be used to merge the different reconstructed datasets into a single coherent one that is closer to the original dataset than any of the simple reconstructed datasets. Finally, we give an instantiation of this reconstruction attack on a decision tree classifier that was learnt using the algorithm C4.5 and evaluate experimentally its efficiency. The results of this experimentation demonstrate that the proposed attack is able to reconstruct a significant part of the original dataset, thus highlighting the need to develop new learning algorithms whose output is specifically tailored to mitigate the success of this type of attack.

[11] Nils Homer et al. "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays". In: *PLOS Genetics* 4.8 (Aug. 2008), pp. 1–9. DOI: 10.1371/journal.pgen.1000167. URL: https://doi.org/10.1371/journal.pgen.1000167.

We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed. In this report we describe a framework for accurately and robustly resolving whether individuals are in a complex genomic DNA mixture using high-density single nucleotide polymorphism (SNP) genotyping microarrays. We develop a theoretical framework for detecting an individual's presence within a mixture, show its limits through simulation, and finally demonstrate experimentally the identification of the presence of genomic DNA of individuals within a series of highly complex genomic mixtures. Our approaches demonstrate straightforward identification of trace amounts ($< 1\%$) of DNA from an individual contributor within a complex mixture. We show how probe-intensity analysis of high-density SNP data can be used, even given the experimental noise of a microarray. We discuss the implications of these findings in two fields: forensics and genome-wide association (GWA) genetic studies. Within forensics, resolving whether an individual is contributing trace amounts of genomic DNA to a complex mixture is a tremendous challenge. Within GWA studies, there is a considerable push to make experimental data publicly available so that the data can be combined with other studies. Our findings show that such an approach does not completely conceal identity, since it is straightforward to assess the probability that a person or relative participated in a GWA study.

[12] Marie-Sarah Lacharité, Brice Minaud, and Kenneth G Paterson. "Improved Reconstruction Attacks on Encrypted Data Using Range Query Leakage". In: *2018 IEEE Symposium on Security and*

Privacy (SP). IEEE. 2018, pp. 297–314. DOI: 10.1109/SP.2018.00002. URL: https://ieeexplore.ieee.org/document/8418610.

We analyse the security of database encryption schemes supporting range queries against persistent adversaries. The bulk of our work applies to a generic setting, where the adversary's view is limited to the set of records matched by each query (known as *access pattern* leakage). We also consider a more specific setting where certain *rank* information is also leaked. The latter is inherent to multiple recent encryption schemes supporting range queries, including Kerschbaum's FH-OPE scheme, Lewi and Wu's order-revealing encryption scheme, and the recently proposed Arx scheme of Poddar et al. We provide three attacks. First, we consider *full reconstruction*, which aims to recover the value of every record, fully negating encryption. We show that for dense datasets, full reconstruction is possible within an expected number of queries $N \log N + O(N)$, where N is the number of distinct plaintext values. This directly improves on a $O(N^2 \log N)$ bound in the same setting by Kellaris et al. (CCS 2016). Second, we present an *approximate reconstruction* attack recovering all plaintext values in a dense dataset within a constant ratio of error, requiring the access pattern leakage of only $O(N)$ queries. Third, we devise an attack in the common setting where the adversary has access to an auxiliary distribution for the target dataset. This third attack proves highly effective on age data from real-world medical data sets. In our experiments, observing only 25 queries was sufficient to reconstruct a majority of records to within 5 years. In combination, our attacks show that current approaches to enabling range queries offer little security when the threat model goes beyond snapshot attacks to include a persistent server-side adversary.

[13] Lingjuan Lyu and Chen Chen. "A Novel Attribute Reconstruction Attack in Federated Learning". In: *arXiv preprint arXiv:2108.06910* (2021). URL: https://doi.org/10.48550/arXiv.2108.06910.

Federated learning (FL) emerged as a promising learning paradigm to enable a multitude of participants to construct a joint ML model without exposing their private training data. Existing FL designs have been shown to exhibit vulnerabilities which can be exploited by adversaries both within and outside of the system to compromise data privacy. However, most current works conduct attacks by leveraging gradients on a small batch of data, which is less practical in FL. In this work, we consider a more practical and interesting scenario in which participants share their epoch-averaged gradients (share gradients after at least 1 epoch of local training) rather than per-example or small batch-averaged gradients as in previous works. We perform the first systematic evaluation of attribute reconstruction attack (ARA) launched by the malicious server in the FL system, and empirically demonstrate that the shared epoch-averaged local model gradients can reveal sensitive attributes of local training data of any victim participant. To achieve this goal, we develop a more effective and efficient gradient matching based method called cos-matching to reconstruct the training data attributes. We evaluate our attacks on a variety of real-world datasets, scenarios, assumptions. Our experiments show that our proposed method achieves better attribute attack performance than most existing baselines.

[14] Guangcan Mai et al. "On the Reconstruction of Face Images from Deep Face Templates". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.5 (2018), pp. 1188–1202. URL: https://ieeexplore.ieee.org/document/8338413.

State-of-the-art face recognition systems are based on deep (convolutional) neural networks. Therefore, it is imperative to determine to what extent face templates derived from deep networks can be inverted to obtain the original face image. In this paper, we study the vulnerabilities of a state-of-the-art face recognition system based on template reconstruction attack. We propose a neighborly de-convolutional neural network (NbNet) to reconstruct face images from their deep templates. In our experiments, we assumed that no knowledge about the target subject and the deep network are available. To train the NbNet reconstruction models,

we augmented two benchmark face datasets (VGG-Face and Multi-PIE) with a large collection of images synthesized using a face generator. The proposed reconstruction was evaluated using type-I (comparing the reconstructed images against the original face images used to generate the deep template) and type-II (comparing the reconstructed images against a different face image of the same subject) attacks. Given the images reconstructed from NbNets, we show that for verification, we achieve TAR of 95.20 percent (58.05 percent) on LFW under type-I (type-II) attacks @ FAR of 0.1 percent. Besides, 96.58 percent (92.84 percent) of the images reconstructed from templates of partition fa (fb) can be identified from partition fa in color FERET. Our study demonstrates the need to secure deep templates in face recognition systems.

[15] Seung Ho Na et al. "Closing the Loophole: Rethinking Reconstruction Attacks in Federated Learning from a Privacy Standpoint". In: *Annual Computer Security Applications Conference*. 2022, pp. 332–345. URL: https://dl.acm.org/doi/abs/10.1145/3564625.3564657.
Federated Learning was deemed as a private distributed learning framework due to the separation of data from the central server. However, recent works have shown that privacy attacks can extract various forms of private information from legacy federated learning. Previous literature describe differential privacy to be effective against membership inference attacks and attribute inference attacks, but our experiments show them to be vulnerable against reconstruction attacks. To understand this outcome, we execute a systematic study of privacy attacks from the standpoint of privacy. The privacy characteristics that reconstruction attacks infringe are different from other privacy attacks, and we suggest that privacy breach occurred at different levels. From our study, reconstruction attack defense methods entail heavy computation or communication costs. To this end, we propose Fragmented Federated Learning (FFL), a lightweight solution against reconstruction attacks. This framework utilizes a simple yet novel gradient obscuring algorithm based on a newly proposed concept called the global gradient and determines which layers are safe for submission to the server. We show empirically in diverse settings that our framework improves practical data privacy of clients in federated learning with an acceptable performance trade-off without increasing communication cost. We aim to provide a new perspective to privacy in federated learning and hope this privacy differentiation can improve future privacy-preserving methods.

[16] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *Proceedings of the 2008 IEEE Symposium on Security and Privacy*. SP '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 111–125. ISBN: 978-0-7695-3168-7. DOI: 10.1109/SP.2008.33. URL: https://ieeexplore.ieee.org/document/4531148.
We present a new class of statistical de- anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge. We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

[17] Salvador Ochoa et al. *Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study*. Tech. rep. MIT, 2001. URL: https://www.researchgate.net/publication/2838440_Reidentification_of_Individuals_in_Chicago's_Homicide_Database_A_Technical_and_Legal_Study.
Many government agencies, hospitals, and other organizations collect personal data of asensitive nature. Often, these groups would like to release their data for statistical analysis bythe scientific community, but

do not want to cause the subjects of the data embarrassment orharassment. To resolve this conflict between privacy and progress, data is often deidentifiedbefore publication. In short, personally identifying information such as names, home addresses,and social security numbers are stripped from the data. We analyzed one such deidentified dataset containing information about Chicago homicide victims over a span of three decades. Bycomparing the records in the Chicago data set with records in the Social Security Death Index,we were able to associate names with, or reidentify, 35% of the victims. This study details thereidentification method and results, and includes a legal review of U.S. regulations related toreidentification. Based on the findings of our project, we recommend removal of these databasesfrom their online locations, and the establishment of national deidentification regulations.

[18] Hyunseok Oh and Youngki Lee. "Exploring Image Reconstruction Attack in Deep Learning Computation Offloading". In: *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*. 2019, pp. 19–24. URL: https://dl.acm.org/doi/10.1145/3325413.3329791.

Deep learning (DL) computation offloading is commonly adopted to enable the use of computation-intensive DL techniques on resource-constrained devices. However, sending private user data to an external server raises a serious privacy concern. In this paper, we introduce a privacy-invading input reconstruction method which utilizes intermediate data of the DL computation pipeline. In doing so, we first define a Peak Signal-to-Noise Ratio (PSNR)-based metric for assessing input reconstruction quality. Then, we simulate a privacy attack on diverse DL models to find out the relationship between DL model structures and performance of privacy attacks. Finally, we provide several insights on DL model structure design to prevent reconstruction-based privacy attacks: using skip-connection, making model deeper, including various DL operations such as inception module.

[19] Rrubaa Panchendrarajan and Suman Bhoi. "Dataset Reconstruction Attack Against Language Models". In: *CEUR Workshop*. 2021. URL: https://ceur-ws.org/Vol-2942/paper1.pdf.

With the advances of deep learning techniques in Natural Language Processing, the last few years have witnessed releases of powerful language models such as BERT and GPT-2. However, applying these general-purpose language models to domain-specific applications requires further fine-tuning using domain-specific private data. Since private data is mostly confidential, information that can be extracted by an adversary with access to the models can lead to serious privacy risks. The majority of privacy attacks on language models infer either targeted information or a few instances from the training dataset. However, inferring the whole training dataset has not been explored in depth which poses far greater risks than disclosure of some instances or partial information of the training data. In this work, we propose a novel data reconstruction attack that also infers the informative words present in the private dataset. Experiment results show that an adversary with black-box query access to a fine-tuned language model can infer the informative words with an accuracy of about 75% and can reconstruct nearly 46.67% of the sentences in the private dataset.

[20] L. Rocher, J.M. Hendrickx, and YA de Montjoye. "Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models". In: *Nature Communications* 10 (2019). DOI: 10.1038/s41467-019-10933-3. URL: https://www.nature.com/articles/s41467-019-10933-3.

While rich medical, behavioral, and socio-demographic data are key to modern data-driven research, their collection and use raise legitimate privacy concerns. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address those concerns. We here propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model,

we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.

[21] Mohammad Al-Rubaie and J Morris Chang. "Reconstruction Attacks Against Mobile-based Continuous Authentication Systems in the Cloud". In: *IEEE Transactions on Information Forensics and Security* 11.12 (2016), pp. 2648–2663. URL: https://ieeexplore.ieee.org/document/7523420.

Continuous authentication for mobile devices using behavioral biometrics is being suggested to complement initial authentication for securing mobile devices, and the cloud services accessed through them. This area has been studied over the past few years, and low error rates were achieved; however, it was based on training and testing using support vector machine (SVM) and other non-privacy-preserving machine learning algorithms. To stress the importance of carefully designed privacy-preserving systems, we investigate the possibility of reconstructing gestures raw data from users' authentication profiles or synthesized samples' testing results. We propose two types of reconstruction attacks based on whether actual user samples are available to the adversary (as in SVM profiles) or not. We also propose two algorithms to reconstruct raw data: a numerical-based algorithm that is specific to one compromised system, and a randomization-based algorithm that can work against almost any compromised system. For our experiments, we selected one compromised and four attacked gesture-based continuous authentication systems from the recent literature. The experiments, performed using a public data set, showed that the attacks were feasible, with a median ranging from 80% to 100% against one attacked system using all types of attacks and algorithms, and a median ranging from 73% to 100% against all attacked systems using the randomization-based algorithm and the negative support vector attack. Finally, we analyze the results, and provide recommendations for building active authentication systems that could resist reconstruction attacks.

[22] Ahmed Salem et al. "Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning". In: *29Th USENIX Security Symposium (USENIX Security 20)*. 2020, pp. 1291–1308. ISBN: 978-1-939133-17-5. URL: https://dl.acm.org/doi/10.5555/3489212.3489285.

Machine learning (ML) has progressed rapidly during the past decade and the major factor that drives such development is the unprecedented large-scale data. As data generation is a continuous process, this leads to ML model owners updating their models frequently with newly-collected data in an online learning scenario. In consequence, if an ML model is queried with the same set of data samples at two different points in time, it will provide different results. In this paper, we investigate whether the change in the output of a black-box ML model before and after being updated can leak information of the dataset used to perform the update, namely the updating set. This constitutes a new attack surface against black-box ML models and such information leakage may compromise the intellectual property and data privacy of the ML model owner. We propose four attacks following an encoder-decoder formulation, which allows inferring diverse information of the updating set. Our new attacks are facilitated by state-of-the-art deep learning techniques. In particular, we propose a hybrid generative model (CBM-GAN) that is based on generative adversarial networks (GANs) but includes a reconstructive loss that allows reconstructing accurate samples. Our experiments show that the proposed attacks achieve strong performance.

[23] Sriram Sankararaman et al. "Genomic Privacy and Limits of Individual Detection in a Pool". In: *Nature Genetics* 41 (), pp. 965–967. URL: https://www.nature.com/articles/ng.436.

Recent studies have demonstrated that statistical methods can be used to detect the presence of a single individual within a study group based on summary data reported from genome-wide association studies (GWAS). We present an analytical and empirical study of the statistical power of such methods. We thereby

aim to provide quantitative guidelines for researchers wishing to make a limited number of SNPs available publicly without compromising subjects' privacy.

[24] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. "Membership Inference Attacks against Machine Learning Models". In: *CoRR* abs/1610.05820 (2016). URL: https://doi.org/10.48550/arXiv.1610.05820.

We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model's training dataset. To perform membership inference against a target model, we make adversarial use of machine learning and train our own inference model to recognize differences in the target model's predictions on the inputs that it trained on versus the inputs that it did not train on. We empirically evaluate our inference techniques on classification models trained by commercial "machine learning as a service" providers such as Google and Amazon. Using realistic datasets and classification tasks, including a hospital discharge dataset whose membership is sensitive from the privacy perspective, we show that these models can be vulnerable to membership inference attacks. We then investigate the factors that influence this leakage and evaluate mitigation strategies.

[25] Pierre Stock et al. "Defending against Reconstruction Attacks with Rényi Differential Privacy". In: *arXiv preprint arXiv:2202.07623* (2022). URL: https://doi.org/10.48550/arXiv.2202.07623.

Reconstruction attacks allow an adversary to regenerate data samples of the training set using access to only a trained model. It has been recently shown that simple heuristics can reconstruct data samples from language models, making this threat scenario an important aspect of model release. Differential privacy is a known solution to such attacks, but is often used with a relatively large privacy budget (epsilon ¿ 8) which does not translate to meaningful guarantees. In this paper we show that, for a same mechanism, we can derive privacy guarantees for reconstruction attacks that are better than the traditional ones from the literature. In particular, we show that larger privacy budgets do not protect against membership inference, but can still protect extraction of rare secrets. We show experimentally that our guarantees hold against various language models, including GPT-2 finetuned on Wikitext-103.

[26] Jiankai Sun et al. "Defending Against Reconstruction Attack in Vertical Federated Learning". In: *arXiv preprint arXiv:2107.09898* (2021). URL: https://doi.org/10.48550/arXiv.2107.09898.

Recently researchers have studied input leakage problems in Federated Learning (FL) where a malicious party can reconstruct sensitive training inputs provided by users from shared gradient. It raises concerns about FL since input leakage contradicts the privacy-preserving intention of using FL. Despite a relatively rich literature on attacks and defenses of input reconstruction in Horizontal FL, input leakage and protection in vertical FL starts to draw researcher's attention recently. In this paper, we study how to defend against input leakage attacks in Vertical FL. We design an adversarial training-based framework that contains three modules: adversarial reconstruction, noise regularization, and distance correlation minimization. Those modules can not only be employed individually but also applied together since they are independent to each other. Through extensive experiments on a large-scale industrial online advertising dataset, we show our framework is effective in protecting input privacy while retaining the model utility.

[27] Latanya Sweeney. *Patient Identifiability in Pharmaceutical Marketing Data*. Tech. rep. Carnegie Mellon University, Harvard University. URL: https://dataprivacylab.org/projects/identifiability/pharma1.pdf.

Does pharmaceutical marketing data expose patient records? In 2003, just after the promulgation of the HIPAA Privacy Rule, a major American pharmaceutical company commissioned a report across 9 states to determine the number of people in those states who may be at risk of being identified if patient pharmacy claims data used for marketing were shared. In May 2003 the report showed that 2.3% of individuals could be uniquely identified from the de-identified prescription records used for marketing purposes at the time and that 6.1% were identifiable to a binsize of 2 (i.e., the record either uniquely related to one named person or related indistinguishably to 2 identified people). These results used prescription information drug, dosage and refill information, patient diagnosis, patient ZIP inferred from pharmacy ZIP, prescription fill date. No explicit patient identifiers (e.g., name or address) appeared in the data. The prescribing doctor was not uniquely identified. Results were based on the states: New York, Illinois, Michigan, Massachusetts, Florida, California, Pennsylvania, Texas, and Arizona. The primary means of re-identification was linking the prescription records to ambulatory and hospital discharge data using patient diagnosis, inferred ZIP, and drug, dosage and refill information to learn more patient demographics and then linking that result to a voter list (or other population register) to learn the names of the subjects of the prescriptions. In comparison, the HIPAA Safe Harbor tends to re-identify about 0.04% of the population, thereby showing that in general more personal information is put at risk in these data than with the HIPAA Safe Harbor, however variability exists in re-identification rates from state to state with some states having re-identification rates less than the HIPAA Safe Harbor. Other privacy observations found in the data, but not part of the analysis, include: (1) the data did not segment or restrict access to special medical classes protected by law, such as psychiatric and HIV related prescriptions; and, (2) the data made it possible to construct a patient's prescription profile over time, which could further increase re-identification risk. This paper summarizes the earlier 2003 report, reviews subsequent publication, and imposes the emergent scientific-legal approach of comparing re-identification rates to the HIPAA Safe Harbor. In the end though, this paper demonstrates the best of measuring de-identification risks while exposing the perils of de-identification as a regime.

[28] Latanya Sweeney. *Simple Demographics Often Identify People Uniquely*. en. Working Paper 3. Carnegie Mellon University, p. 34. URL: https://dataprivacylab.org/projects/identifiability/paper1.pdf.

In this document, I report on experiments I conducted using 1990 U.S. Census summary data to determine how many individuals within geographically situated populations had combinations of demographic values that occurred infrequently. It was found that combinations of few characteristics often combine in populations to uniquely or nearly uniquely identify some individuals. Clearly, data released containing such information about these individuals should not be considered anonymous. Yet, health and other person-specific data are publicly available in this form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only place, gender, date of birth, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

[29] Shangyu Xie and Yuan Hong. "Reconstruction Attack on Instance Encoding for Language Understanding". In: *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 2021. DOI: 10.18653/v1/2021.emnlp-main.154. URL: https://aclanthology.org/2021.emnlp-main.154.

A private learning scheme TextHide was recently proposed to protect the private text data during the training phase via so-called instance encoding. We propose a novel reconstruction attack to break TextHide by recovering the private training data, and thus unveil the privacy risks of instance encoding. We have experimentally validated the effectiveness of the reconstruction attack with two commonly-used datasets for sentence classification. Our attack would advance the development of privacy preserving machine learning in the context of natural language processing.

[30] Peng Ye et al. "Feature Reconstruction Attacks and Countermeasures of DNN training in Vertical Federated Learning". In: *arXiv preprint arXiv:2210.06771* (2022). URL: https://doi.org/10.48550/arXiv.2210.06771.

Federated learning (FL) has increasingly been deployed, in its vertical form, among organizations to facilitate secure collaborative training over siloed data. In vertical FL (VFL), participants hold disjoint features of the same set of sample instances. Among them, only one has labels. This participant, known as the active party, initiates the training and interacts with the other participants, known as the passive parties. Despite the increasing adoption of VFL, it remains largely unknown if and how the active party can extract feature data from the passive party, especially when training deep neural network (DNN) models. This paper makes the first attempt to study the feature security problem of DNN training in VFL. We consider a DNN model partitioned between active and passive parties, where the latter only holds a subset of the input layer and exhibits some categorical features of binary values. Using a reduction from the Exact Cover problem, we prove that reconstructing those binary features is NP-hard. Through analysis, we demonstrate that, unless the feature dimension is exceedingly large, it remains feasible, both theoretically and practically, to launch a reconstruction attack with an efficient search-based algorithm that prevails over current feature protection techniques. To address this problem, we develop a novel feature protection scheme against the reconstruction attack that effectively misleads the search to some pre-specified random values. With an extensive set of experiments, we show that our protection scheme sustains the feature reconstruction attack in various VFL applications at no expense of accuracy loss.

2 Papers Outlining Theory for an Attack

Our second category of papers focuses on work that does not empirically demonstrate an attack, but lays out the theoretical foundations necessary to carry an attack out. This category is broader than the other three categories, because it is difficult to precisely limit the areas of mathematics necessary for carrying out a privacy-violating attack. Many areas of mathematics are, in general, relevant. We have used our judgement in selecting the papers represented here, aiming to cover this literature with reasonable thoroughness, but without including work that would require considerable effort to relate back to immediate privacy-violating attacks.

Theoretical Foundations

[31] Emmanuel J Candes and Terence Tao. "Decoding by Linear Programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215. ISSN: 1557-9654. DOI: 10.1109/TIT.2005.858979. URL: https://ieeexplore.ieee.org/document/1542412.

This paper considers a natural error correcting problem with real valued input/output. We wish to recover an input vector $f \in \mathbb{R}^n$ from corrupted measurements $y = Af + e$. Here, A is an m by n (coding) matrix and e is an arbitrary and unknown vector of errors. Is it possible to recover f exactly from the data y?

We prove that under suitable conditions on the coding matrix A, the input f is the unique solution to the

ℓ_1-minimization problem ($||x||_{\ell_1} := \sum_i |x_i|$)

$$\min_{g \in \mathbb{R}_n} ||y - Ag||_{\ell_1}$$

provided that the support of the vector of errors is not too large, $||e||_{\ell_0} := |\{i : e_i \neq 0\}| \leq \rho \cdot m$ for some $\rho > 0$. In short, f can be recovered exactly by solving a simple convex optimization problem (which one can recast as a linear program). In addition, numerical experiments suggest that this recovery procedure works unreasonably well; f is recovered exactly even in situations where a significant fraction of the output is corrupted.

This work is related to the problem of finding sparse solutions to vastly underdetermined systems of linear equations. There are also significant connections with the problem of recovering signals from highly incomplete measurements. In fact, the results introduced in this paper improve on our earlier work [Candès & Tao 2006 Near Optimal Signal Recovery]. Finally, underlying the success of ℓ_1 is a crucial property we call the uniform uncertainty principle that we shall describe in detail.

[32] Emmanuel Candès et al. "Error Correction via Linear Programming". In: *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*. 2005, pp. 668–681. DOI: 10.1109/SFCS.2005.5464411. URL: https://ieeexplore.ieee.org/document/5464411.

Suppose we wish to transmit a vector $f \in \mathbb{R}_n$ reliably. A frequently discussed approach consists in encoding f with an m by n coding matrix A. Assume now that a fraction of the entries of Af are corrupted in a completely arbitrary fashion by an error e. We do not know which entries are affected nor do we know how they are affected. Is it possible to recover f exactly from the corrupted m-dimensional vector $y = Af + e$?

[33] Benny Chor et al. *Tracing Traitors*. 1994. URL: http://web.cs.ucla.edu/~miodrag/cs259-security/chor94tracing.pdf.

We give cryptographic schemes that help trace the source of leaks when sensitive or proprietary data is made available to a large set of parties. A very relevant application is in the context of pay television, where only paying customers should be able to view certain programs. In this application, the programs are normally encrypted, and then the sensitive data is the decryption keys that are given to paying customers. If a pirate decoder is found, it is desirable to reveal the source of its decryption keys. We describe fully resilient schemes which can be used against any decoder which decrypts with nonnegligible probability. Since there is typically little demand for decoders which decrypt only a small fraction of the transmissions (even if it is nonnegligible), we further introduce threshold tracing schemes which can only be used against decoders which succeed in decryption with probability greater than some threshold. Threshold schemes are considerably more efficient than fully resilient schemes.

[34] Anindya De. "Lower Bounds in Differential Privacy". In: *Theory of Cryptography Conference*. Ed. by Ronald Cramer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 321–338. ISBN: 978-3-642-28914-9. DOI: 10.1007/978-3-642-28914-9_18. URL: https://link.springer.com/chapter/10.1007/978-3-642-28914-9_18.

This paper is about private data analysis, in which a trusted curator holding a confidential database responds to real vector-valued queries. A common approach to ensuring privacy for the database elements is to add appropriately generated random noise to the answers, releasing only these noisy responses. A line of study initiated in [35] examines the amount of distortion needed to prevent privacy violations of various kinds. The results in the literature vary according to several parameters, including the size of the database, the size of the universe from which data elements are drawn, the "amount" of privacy desired, and for the purposes of the current work, the arity of the query. In this paper we sharpen and unify these bounds. Our foremost result combines the techniques of Hardt and Talwar and McGregor et al. to obtain linear lower bounds on distortion when providing differential privacy for a (contrived) class of low-sensitivity queries. (A query has

low sensitivity if the data of a single individual has small effect on the answer.) Several structural results follow as immediate corollaries:

- We separate so-called *counting* queries from arbitrary *low-sensitivity* queries, proving the latter requires more noise, or distortion, than does the former;
- We separate $(\varepsilon, 0)$-differential privacy from its well-studied relaxation (ε, δ)-differential privacy, even when $\delta \in 2^{-o(n)}$ is negligible in the size n of the database, proving the latter requires less distortion than the former;
- We demonstrate that (ε, δ)-differential privacy is much weaker than $(\varepsilon, 0)$-differential privacy in terms of mutual information of the transcript of the mechanism with the database, even when $\delta \in 2^{-o(n)}$ is negligible in the size n of the database.

We also simplify the lower bounds on noise for counting queries in Hardt and Talwar and also make them unconditional. Further, we use a characterization of (ε, δ)-differential privacy from McGregor et al. to obtain lower bounds on the distortion needed to ensure (ε, δ)-differential privacy for $\varepsilon, \delta > 0$. We next revisit the LP decoding argument of [36] and combine it with a recent result of Rudelson to improve on a result of [39] on noise lower bounds for privately releasing ℓ-way marginals.

[35] Irit Dinur and Kobbi Nissim. "Revealing Information While Preserving Privacy". In: *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '03. San Diego, California: ACM, 2003, pp. 202–210. ISBN: 1-58113-670-6. DOI: 10.1145/773153.773173. URL: http://doi.acm.org/10.1145/773153.773173.
We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an n-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to be answered by $\Sigma_{i \varepsilon q} d_i$. Our main result is a polynomial reconstruction algorithm of data from noisy (perturbed) subset sums. Applying this reconstruction algorithm to statistical databases we show that in order to achieve privacy one has to add perturbation of magnitude $\Omega(\sqrt{n})$. That is, smaller perturbation always results in a strong violation of privacy. We show that this result is tight by exemplifying access algorithms for statistical databases that preserve privacy while adding perturbation of magnitude $\tilde{O}(\sqrt{n})$. For time-T bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is $\approx \sqrt{T}$.

[36] Cynthia Dwork, Frank McSherry, and Kunal Talwar. "The Price of Privacy and the Limits of LP Decoding". In: *Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing STOC '07*. ACM Digital Library, 2007, pp. 85–94. DOI: 10.1145/1250790.1250804. URL: https://dl.acm.org/doi/10.1145/1250790.1250804.
This work is at the intersection of two lines of research. One line, initiated by Dinur and Nissim, investigates the price, in accuracy, of protecting privacy in a statistical database. The second, growing from an extensive literature on compressed sensing (see in particular the work of Donoho and collaborators [Donoho & Johnstone, Chen et al, Donoho & Huo, Donoho]) and explicitly connected to error-correcting codes by Candès and Tao[31] (see also Candès, Rudelson, Tao, and Vershyin[32]), is in the use of linear programming for error correction. Our principal result is the discovery of a sharp threshhold $\rho^* \approx 0.239$, so that if $\rho < \rho^*$ and A is a random $m \times n$ encoding matrix of independently chosen standard Gaussians, where $m = O(n)$, then with overwhelming probability over choice of A, for all $x \in \mathbb{R}^n$, LP decoding corrects $\lfloor \rho m \rfloor$ arbitrary errors in the encoding Ax, while decoding can be made to fail if the error rate exceeds ρ^*. Our bound resolves an open question of Candès, Rudelson, Tao, and Vershyin[32] and (oddly, but explicably) refutes empirical conclusions of Donoho and Candès et al[31]. By scaling and rounding we can easily transform these results to obtain polynomial-time decodable random linear codes with polynomial-sized alphabets tolerating any $\rho < \rho^* \approx 0.239$ fraction of arbitrary errors. In the context of privacy-preserving datamining our results say that any privacy mechanism, interactive or non-interactive, providing reasonably accurate answers to a 0.761

fraction of randomly generated weighted subset sum queries, and arbitrary answers on the remaining 0.239 fraction, is blatantly non-private.

[37] Cynthia Dwork and Moni Naor. "On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy". In: *Journal of Privacy and Confidentiality* 2.1 (2010), pp. 93–107. DOI: 10.29012/jpc.v2i1.585. URL: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/585.
In 1977 Tore Dalenius articulated a desideratum for statistical databases: nothing about an individual should be learnable from the database that cannot be learned without access to the database. We give a general impossibility result showing that a natural formalization of Dalenius goal cannot be achieved if the database is useful. The key obstacle is the side information that may be available to an adversary. Our results hold under very general conditions regarding the database, the notion of privacy violation, and the notion of utility. Contrary to intuition, a variant of the result threatens the privacy even of someone not in the database. This state of affairs motivated the notion of differential privacy, a strong *ad omnia* privacy which, intuitively, captures the increased risk to ones privacy incurred by participating in a database.

[38] Cynthia Dwork and Sergey Yekhanin. *New Efficient Attacks on Statistical Disclosure Control Mechanisms*. 2008. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2008/08/dy08.pdf.
The goal of a statistical database is to provide statistics about a population while simultaneously protecting the privacy of the individual records in the database. The tension between privacy and usability of statistical databases has attracted much attention in statistics, theoretical computer science, security, and database communities in recent years. A line of research initiated by Dinur and Nissim investigates for a particular type of queries, lower bounds on the distortion needed in order to prevent gross violations of privacy. The first result in the current paper simplifies and sharpens the Dinur and Nissim result. The Dinur-Nissim style results are strong because they demonstrate insecurity of all low-distortion privacy mechanisms. The attacks have an all-or-nothing flavor: letting n denote the size of the database, $\Omega(n)$ queries are made before anything is learned, at which point $\Theta(n)$ secret bits are revealed. Restricting attention to a wide and realistic subset of possible low-distortion mechanisms, our second result is a more acute attack, requiring only a fixed number of queries for each bit revealed.

[39] Shiva Kasiviswanathan et al. "The Price of Privately Releasing Contingency Tables and the Spectra of Random Matrices With Correlated Rows". In: *Proceedings of the 42nd ACM Symposium on the Theory of Computing*. 2010, pp. 775–784. DOI: 10.1145/1806689.1806795. URL: http://www-personal.umich.edu/~rudelson/papers/krsu-privacy.pdf.
Marginal (contingency) tables are the method of choice for government agencies releasing statistical summaries of categorical data. In this paper, we consider lower bounds on how much distortion (noise) is necessary in these tables to provide privacy guarantees when the data being summarized is sensitive. We extend a line of recent work on lower bounds on noise for private data analysis [35, 36, 38], [Dwork, McSherry, Nissim, Smith TCC'06] to a natural and important class of functionalities. Our investigation also leads to new results on the spectra of random matrices with correlated rows. Consider a database D consisting of n rows (one per individual), each row comprising d binary attributes. For any subset of T attributes of size $|T| = k$, the marginal table for T has 2^k entries; each entry counts how many times in the database a particular setting of these attributes occurs. We provide lower bounds for releasing k-attribute marginal tables under (i) *minimal privacy*, a general privacy notion which captures a large class of privacy definitions, and (ii) *differential privacy*, a rigorous notion of privacy that has received extensive recent study. Our main contributions are:

- We give efficient polynomial time attacks which allow an adversary to reconstruct sensitive information given insufficiently perturbed marginal table releases. Using these reconstruction attacks, we show that for releasing all k-attribute marginal tables with constant k, $\tilde{\Omega}(\min\{\sqrt{n}, \sqrt{d^{k-1}}\})$ average distortion per entry is necessary for any privacy notion satisfying at least a minimalistic privacy guarantee. Under this privacy guarantee this bound is tight.
- Our above reconstruction-based attacks require a new lower bound on the least singular value of a random matrix with correlated rows. For a constant k, consider a matrix $M^{(k)}$ with d^k rows which are formed by taking all possible k-way entry-wise products of an underlying set of d random vectors from $\{0,1\}^n$. We show that even if $M^{(k)}$ is nearly square its least singular value is $\tilde{\Omega}(\sqrt{d^k})$ with high probability — asymptotically, the same bound as one gets for a matrix with *independent* rows. The proof introduces several new tools for dealing with random matrices with correlated entries and could be of independent interest.
- We obtain stronger lower bounds for differential privacy. For releasing all k-attribute marginal tables with constant k, previous work showed that $\tilde{O}(\min\{n, (n^2 d)^{1/3}, \sqrt{d^k}\})$ average distortion per entry is *sufficient* for satisfying differential privacy (ignoring the dependence on privacy parameters). We give a lower bound of $\tilde{\Omega}(\min\{\sqrt{n}, \sqrt{d^k}\})$, which is tight for $n = \tilde{\Omega}(d^k)$. Moreover, for a natural and popular class of mechanisms based on adding instance-independent noise, our lower bound can be strengthened to $\tilde{\Omega}(\sqrt{d^k})$, which is tight for all n. Our lower bounds for differential privacy extend even to non-constant k, losing roughly a factor of $\sqrt{2^k}$ compared to best-known upper bounds for large n.

.

[40] Shiva Prasad Kasiviswanathan, Mark Rudelson, and Adam Smith. "The Power of Linear Reconstruction Attacks". In: *Proceedings of the Twenty-fourth Annual ACM-SIAM Symposium on Discrete Algorithms SODA '13*. ACM Digital Library, 2013, pp. 1415–1433. URL: https://doi.org/10.48550/arXiv.1210.2381.

We consider the power of linear reconstruction attacks in statistical data privacy, showing that they can be applied to a much wider range of settings than previously understood. Linear attacks have been studied before (Dinur and Nissim PODS'03[35], Dwork, McSherry and Talwar STOC'07[36], Kasiviswanathan, Rudelson, Smith and Ullman STOC'10[39], De TCC'12[34], Muthukrishnan and Nikolov STOC'12) but have so far been applied only in settings with releases that are obviously linear. Consider a database curator who manages a database of sensitive information but wants to release statistics about how a sensitive attribute (say, disease) in the database relates to some nonsensitive attributes (e.g., postal code, age, gender, etc). We show one can mount linear reconstruction attacks based on any release that gives: a) the fraction of records that satisfy a given non-degenerate boolean function. Such releases include contingency tables (previously studied by Kasiviswanathan, Rudelson, Smith and Ullman STOC'10) as well as more complex outputs like the error rate of classifiers such as decision trees; b) any one of a large class of M-estimators (that is, the output of empirical risk minimization algorithms), including the standard estimators for linear and logistic regression. We make two contributions: first, we show how these types of releases can be transformed into a linear format, making them amenable to existing polynomial-time reconstruction algorithms. This is already perhaps surprising, since many of the above releases (like M-estimators) are obtained by solving highly nonlinear formulations. Second, we show how to analyze the resulting attacks under various distributional assumptions on the data. Specifically, we consider a setting in which the same statistic (either a) or b) above) is released about how the sensitive attribute relates to all subsets of size k (out of a total of d) nonsensitive boolean attributes.

[41] Prottay Protivash et al. *Reconstruction Attacks on Aggressive Relaxations of Differential Privacy*. 2022. DOI: 10.48550/ARXIV.2209.03905. URL: https://doi.org/10.48550/arXiv.2209.03905.

Differential privacy is a widely accepted formal privacy definition that allows aggregate information about a dataset to be released while controlling privacy leakage for individuals whose records appear in the data. Due to the unavoidable tension between privacy and utility, there have been many works trying to relax the requirements of differential privacy to achieve greater utility. One class of relaxation, which is starting to gain support outside the privacy community is embodied by the definitions of individual differential privacy (IDP) and bootstrap differential privacy (BDP). The original version of differential privacy defines a set of neighboring database pairs and achieves its privacy guarantees by requiring that each pair of neighbors should be nearly indistinguishable to an attacker. The privacy definitions we study, however, aggressively reduce the set of neighboring pairs that are protected. Both IDP and BDP define a measure of ”privacy loss” that satisfies formal privacy properties such as postprocessing invariance and composition, and achieve dramatically better utility than the traditional variants of differential privacy. However, there is a significant downside - we show that they allow a significant portion of the dataset to be reconstructed using algorithms that have arbitrarily low privacy loss under their privacy accounting rules. We demonstrate these attacks using the preferred mechanisms of these privacy definitions. In particular, we design a set of queries that, when protected by these mechanisms with high noise settings (i.e., with claims of very low privacy loss), yield more precise information about the dataset than if they were not protected at all.

[42] Novi Quadrianto et al. “Estimating Labels from Label Proportions”. In: *Journal of Machine Learning Research* (2009), pp. 2349–2374. DOI: 10.1145/1390156.1390254. URL: https://jmlr.csail.mit.edu/papers/volume10/quadrianto09a/quadrianto09a.pdf.
Consider the following problem: given sets of unlabeled observations, each set with known labelproportions, predict the labels of another set of observations, possibly with known label proportions. This problem occurs in areas like e-commerce, politics, spam filtering and improper contentdetection. We present consistent estimators which can reconstruct the correct labels with high probability in a uniform convergence sense. Experiments show that our method works well in practice.

3 Papers Focused on Standards and Pedagogy

Our third category of papers focuses on a small, growing body of work aimed at education, outreach, and the establishment of standards relevant to privacy-violating attacks.

Standards and Pedagogy

[43] Simson Garfinkel. *De-Identification of Personal Information*. Internal Report 8053. National Institute of Standards and Technology, Oct. 2015. URL: https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf.
De-identification removes identifying information from a dataset so that individual data cannotvbe linked with specific individuals. De-identification can reduce the privacy risk associated with collecting, processing, archiving, distributing or publishing information. De-identification thus attempts to balance the contradictory goals of using and sharing personal information while protecting privacy. Several U.S laws, regulations and policies specify that data should be deidentified prior to sharing. In recent years researchers have shown that some de-identified data can sometimes be re-identified. Many different kinds of information can be de-identified, including structured information, free format text, multimedia, and medical imagery. This document summarizes roughly two decades of de-identification research, discusses current practices, and presents opportunities for future research.

[44] Simson Garfinkel, J.M. Abowd, and Christian Martindale. "Understanding Database Reconstruction Attacks on Public Data". In: *acmqueue* 16.5 (2018). URL: https://queue.acm.org/detail.cfm?id=3295691.
No abstract. Pedagogical; walks through toy examples of reconstruction attacks.

4 Papers with Background on the 2020 Census DAS

Our fourth category of papers collects a small number of papers necessary for understanding the design and implementation of the 2020 Census of Population and Housing Disclosure Avoidance System, the construction of which was heavily inspired by awareness of attacks described in papers that appear in other categories of this bibliography.

Background on the 2020 Census DAS

[45] John Abowd et al. *Geographic Spines in the 2020 Census Disclosure Avoidance System TopDown Algorithm*. 2022. URL: https://doi.org/10.48550/arXiv.2203.16654.
The TopDown Algorithm (TDA) first produces differentially private counts at the nation and then produces counts at lower geolevels (e.g.: state, county, etc.) subject to the constraint that all query answers in lower geolevels are consistent with those at previously estimated geolevels. This paper describes the three sets of definitions of these geolevels, or the geographic spines, that are implemented within TDA. These include the standard Census geographic spine and two other spines that improve accuracy in geographic areas that are far from the standard Census spine, such as cities, towns, and/or AIAN areas. The third such spine, which is called the optimized spine, also modifies the privacy-loss budget allocated to the entities within the geolevels, or the geounits, to ensure the privacy-loss budget is used efficiently within TDA.

[46] John Abowd et al. *The 2020 Census Disclosure Avoidance System TopDown Algorithm*. 2022. URL: https://hdsr.mitpress.mit.edu/pub/7evz361i/release/1.
The Census TopDown Algorithm (TDA) is a disclosure avoidance system using differential privacy for privacy-loss accounting. The algorithm ingests the final, edited version of the 2020 Census data and the final tabulation geographic definitions. The algorithm then creates noisy versions of key queries on the data, referred to as measurements, using zero-Concentrated Differential Privacy. Another key aspect of the TDA are invariants, statistics that the Census Bureau has determined, as matter of policy, to exclude from the privacy-loss accounting. The TDA postprocesses the measurements together with the invariants to produce a Microdata Detail File (MDF) that contains one record for each person and one record for each housing unit enumerated in the 2020 Census. The MDF is passed to the 2020 Census tabulation system to produce the 2020 Census Redistricting Data (P.L. 94-171) Summary File. This article describes the mathematics and testing of the TDA for this purpose.

[47] Mark Bun and Thomas Steinke. "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds". In: *CoRR* abs/1605.02065 (2016). URL: https://doi.org/10.48550/arXiv.1605.02065.
"Concentrated differential privacy" was recently introduced by Dwork and Rothblum as a relaxation of differential privacy, which permits sharper analyses of many privacy-preserving computations. We present an alternative formulation of the concept of concentrated differential privacy in terms of the Renyi divergence between the distributions obtained by running an algorithm on neighboring inputs. With this reformulation in hand, we prove sharper quantitative results, establish lower bounds, and raise a few new questions. We also unify this approach with approximate differential privacy by giving an appropriate definition of "approximate concentrated differential privacy."

[48] Clement Canonne, Gautam Kamath, and Thomas Steinke. *The Discrete Gaussian for Differential Privacy*. 2021. DOI: 10.48550/arXiv.2004.00010. URL: https://doi.org/10.48550/arXiv.2004.00010.

A key tool for building differentially private systems is adding Gaussian noise to the output of a function evaluated on a sensitive dataset. Unfortunately, using a continuous distribution presents several practical challenges. First and foremost, finite computers cannot exactly represent samples from continuous distributions, and previous work has demonstrated that seemingly innocuous numerical errors can entirely destroy privacy. Moreover, when the underlying data is itself discrete (e.g., population counts), adding continuous noise makes the result less interpretable. With these shortcomings in mind, we introduce and analyze the discrete Gaussian in the context of differential privacy. Specifically, we theoretically and experimentally show that adding discrete Gaussian noise provides essentially the same privacy and accuracy guarantees as the addition of continuous Gaussian noise. We also present an simple and efficient algorithm for exact sampling from this distribution. This demonstrates its applicability for privately answering counting queries, or more generally, low-sensitivity integer-valued queries.

5 Papers Studying Membership Attacks

Our fifth and final category of papers studies attacks that attempt a very limited form of privacy-violating inference: inferring whether a target person's data was present in the training of a model or publication of a set of data. In Census Bureau and Internal Revenue Service terminology, these papers look whether "fact-of-filing" information was unintentionally disclosed. Though we think of these papers as a subset of the category 1 papers, there are sufficiently many membership attack papers to justify listing them separately.

Membership Attacks

[49] Michael Backes et al. "Membership Privacy in MicroRNA-based Studies". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016, pp. 319–330. DOI: 10.1145/2976749.2978355. URL: https://dl.acm.org/doi/10.1145/2976749.2978355.

The continuous decrease in cost of molecular profiling tests is revolutionizing medical research and practice, but it also raises new privacy concerns. One of the first attacks against privacy of biological data, proposed by Homer et al. in 2008, showed that, by knowing parts of the genome of a given individual and summary statistics of a genome-based study, it is possible to detect if this individual participated in the study. Since then, a lot of work has been carried out to further study the theoretical limits and to counter the genome-based membership inference attack. However, genomic data are by no means the only or the most influential biological data threatening personal privacy. For instance, whereas the genome informs us about the risk of developing some diseases in the future, epigenetic biomarkers, such as microRNAs, are directly and deterministically affected by our health condition including most common severe diseases. In this paper, we show that the membership inference attack also threatens the privacy of individuals contributing their microRNA expressions to scientific studies. Our results on real and public microRNA expression data demonstrate that disease-specific datasets are especially prone to membership detection, offering a true-positive rate of up to 77% at a false-negative rate of less than 1%. We present two attacks: one relying on the L_1 distance and the other based on the likelihood-ratio test. We show that the likelihood-ratio test provides the highest adversarial success and we derive a theoretical limit on this success. In order to mitigate the membership inference, we propose and evaluate both a differentially private mechanism and a hiding mechanism. We also consider two types of adversarial prior knowledge for the differentially private

mechanism and show that, for relatively large datasets, this mechanism can protect the privacy of participants in miRNA-based studies against strong adversaries without degrading the data utility too much. Based on our findings and given the current number of miRNAs, we recommend to only release summary statistics of datasets containing at least a couple of hundred individuals.

[50] Nicholas Carlini et al. "Membership Inference Attacks from First Principles". In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 1897–1914. URL: https://doi.org/10.48550/arXiv.2112.03570.
A membership inference attack allows an adversary to query a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset. These attacks are currently evaluated using average-case "accuracy" metrics that fail to characterize whether the attack can confidently identify any members of the training set. We argue that attacks should instead be evaluated by computing their true-positive rate at low (e.g., $\leq 0.1\%$) false-positive rates, and find most prior attacks perform poorly when evaluated in this way. To address this we develop a Likelihood Ratio Attack (LiRA) that carefully combines multiple ideas from the literature. Our attack is 10× more powerful at low false-positive rates, and also strictly dominates prior attacks on existing metrics.

[51] Dingfan Chen et al. "GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models". In: *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 2020, pp. 343–362. DOI: 10.1145/3372297.3417238. URL: https://dl.acm.org/doi/10.1145/3372297.3417238.
Deep learning has achieved overwhelming success, spanning from discriminative models to generative models. In particular, deep generative models have facilitated a new level of performance in a myriad of areas, ranging from media manipulation to sanitized dataset generation. Despite the great success, the potential risks of privacy breach caused by generative models have not been analyzed systematically. In this paper, we focus on membership inference attack against deep generative models that reveals information about the training data used for victim models. Specifically, we present the first taxonomy of membership inference attacks, encompassing not only existing attacks but also our novel ones. In addition, we propose the first generic attack model that can be instantiated in a large range of settings and is applicable to various kinds of deep generative models. Moreover, we provide a theoretically grounded attack calibration technique, which consistently boosts the attack performance in all cases, across different attack settings, data modalities, and training configurations. We complement the systematic analysis of attack performance by a comprehensive experimental study, that investigates the effectiveness of various attacks w.r.t. model type and training configurations, over three diverse application scenarios (i.e., images, medical data, and location data).

[52] Christopher A. Choquette-Choo et al. "Label-Only Membership Inference Attacks". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 1964–1974. URL: https://proceedings.mlr.press/v139/choquette-choo21a.html.
Membership inference is one of the simplest privacy threats faced by machine learning models that are trained on private sensitive data. In this attack, an adversary infers whether a particular point was used to train the model, or not, by observing the model's predictions. Whereas current attack methods all require access to the model's predicted confidence score, we introduce a label-only attack that instead evaluates the robustness of the model's predicted (hard) labels under perturbations of the input, to infer membership. Our label-only attack is not only as-effective as attacks requiring access to confidence scores, it also demonstrates that a class of defenses against membership inference, which we call "confidence masking" because they obfuscate the confidence scores to thwart attacks, are insufficient to prevent the leakage of private information. Our experiments show that training with differential privacy or strong L2 regularization are the only current

defenses that meaningfully decrease leakage of private information, even for points that are outliers of the training distribution.

[53] Jamie Hayes et al. "Logan: Membership Inference Attacks Against Generative Models". In: *arXiv preprint arXiv:1705.07663* (2017). DOI: 10.48550/arXiv.1705.07663. URL: https://arxiv.org/abs/1705.07663.
Generative models estimate the underlying distribution of a dataset to generate realistic samples according to that distribution. In this paper, we present the first membership inference attacks against generative models: given a data point, the adversary determines whether or not it was used to train the model. Our attacks leverage Generative Adversarial Networks (GANs), which combine a discriminative and a generative model, to detect overfitting and recognize inputs that were part of training datasets, using the discriminator's capacity to learn statistical differences in distributions. We present attacks based on both white-box and black-box access to the target model, against several state-of-the-art generative models, over datasets of complex representations of faces (LFW), objects (CIFAR-10), and medical images (Diabetic Retinopathy). We also discuss the sensitivity of the attacks to different training parameters, and their robustness against mitigation strategies, finding that defenses are either ineffective or lead to significantly worse performances of the generative models in terms of training stability and/or sample quality.

[54] Xinlei He et al. "Node-level Membership Inference Attacks Against Graph Neural Networks". In: *arXiv preprint arXiv:2102.05429* (2021). DOI: 10.48550/arXiv.2102.05429. URL: https://arXiv.org/abs/2102.05429.
Many real-world data comes in the form of graphs, such as social networks and protein structure. To fully utilize the information contained in graph data, a new family of machine learning (ML) models, namely graph neural networks (GNNs), has been introduced. Previous studies have shown that machine learning models are vulnerable to privacy attacks. However, most of the current efforts concentrate on ML models trained on data from the Euclidean space, like images and texts. On the other hand, privacy risks stemming from GNNs remain largely unstudied. In this paper, we fill the gap by performing the first comprehensive analysis of node-level membership inference attacks against GNNs. We systematically define the threat models and propose three node-level membership inference attacks based on an adversary's background knowledge. Our evaluation on three GNN structures and four benchmark datasets shows that GNNs are vulnerable to node-level membership inference even when the adversary has minimal background knowledge. Besides, we show that graph density and feature similarity have a major impact on the attack's success. We further investigate two defense mechanisms and the empirical results indicate that these defenses can reduce the attack performance but with moderate utility loss.

[55] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models." In: *Proc. Priv. Enhancing Technol.* 2019.4 (2019), pp. 232–249. DOI: 10.2478/popets-2019-0067. URL: https://petsymposium.org/popets/2019/popets-2019-0067.php.
We present two information leakage attacks that outperform previous work on membership inference against generative models. The first attack allows membership inference without assumptions on the type of the generative model. Contrary to previous evaluation metrics for generative models, like Kernel Density Estimation, it only considers samples of the model which are close to training data records. The second attack specifically targets Variational Autoencoders, achieving high membership inference accuracy. Furthermore, previous work mostly considers membership inference adversaries who perform single record membership inference. We argue for considering regulatory actors who perform set membership inference to identify the use of specific datasets for training. The attacks are evaluated on two generative model architectures, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), trained on standard image

datasets. Our results show that the two attacks yield success rates superior to previous work on most data sets while at the same time having only very mild assumptions. We envision the two attacks in combination with the membership inference attack type formalization as especially useful. For example, to enforce data privacy standards and automatically assessing model quality in machine learning as a service setups. In practice, our work motivates the use of GANs since they prove less vulnerable against information leakage attacks while producing detailed samples.

[56] Hongsheng Hu et al. "Membership Inference Attacks on Machine Learning: a Survey". In: *ACM Computing Surveys (CSUR)* 54.11s (2022), pp. 1–37. DOI: 10.1145/3523273. URL: https://dl.acm.org/doi/10.1145/3523273.

Machine learning (ML) models have been widely applied to various applications, including image classification, text generation, audio recognition, and graph data analysis. However, recent studies have shown that ML models are vulnerable to membership inference attacks (MIAs), which aim to infer whether a data record was used to train a target model or not. MIAs on ML models can directly lead to a privacy breach. For example, via identifying the fact that a clinical record that has been used to train a model associated with a certain disease, an attacker can infer that the owner of the clinical record has the disease with a high chance. In recent years, MIAs have been shown to be effective on various ML models, e.g., classification models and generative models. Meanwhile, many defense methods have been proposed to mitigate MIAs. Although MIAs on ML models form a newly emerging and rapidly growing research area, there has been no systematic survey on this topic yet. In this paper, we conduct the first comprehensive survey on membership inference attacks and defenses. We provide the taxonomies for both attacks and defenses, based on their characterizations, and discuss their pros and cons. Based on the limitations and gaps identified in this survey, we point out several promising future research directions to inspire the researchers who wish to follow this area. This survey not only serves as a reference for the research community but also provides a clear description for researchers outside this research domain. To further help the researchers, we have created an online resource repository, which we will keep updated with future relevant work. Interested readers can find the repository at this https URL.

[57] Bo Hui et al. "Practical Blind Membership Inference Attack via Differential Comparisons". In: *arXiv preprint arXiv:2101.01341* (2021). DOI: 10.48550/arXiv.2101.01341. URL: https://arxiv.org/abs/2101.01341.

Membership inference (MI) attacks affect user privacy by inferring whether given data samples have been used to train a target learning model, e.g., a deep neural network. There are two types of MI attacks in the literature, i.e., these with and without shadow models. The success of the former heavily depends on the quality of the shadow model, i.e., the transferability between the shadow and the target; the latter, given only blackbox probing access to the target model, cannot make an effective inference of unknowns, compared with MI attacks using shadow models, due to the insufficient number of qualified samples labeled with ground truth membership information. In this paper, we propose an MI attack, called BlindMI, which probes the target model and extracts membership semantics via a novel approach, called differential comparison. The high-level idea is that BlindMI first generates a dataset with nonmembers via transforming existing samples into new samples, and then differentially moves samples from a target dataset to the generated, non-member set in an iterative manner. If the differential move of a sample increases the set distance, BlindMI considers the sample as non-member and vice versa. BlindMI was evaluated by comparing it with state-of-the-art MI attack algorithms. Our evaluation shows that BlindMI improves F1-score by nearly 20% when compared to state-of-the-art on some datasets, such as Purchase-50 and Birds-200, in the blind setting where the adversary does not know the target model's architecture and the target dataset's ground truth labels. We also show that BlindMI can defeat state-of-the-art defenses.

[58] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. "Membership Inference Attack Susceptibility of Clinical Language Models". In: *arXiv preprint arXiv:2104.08305* (2021). DOI: 10.48550/arXiv.2104.08305. URL: https://arxiv.org/abs/2104.08305.
Deep Neural Network (DNN) models have been shown to have high empirical privacy leakages. Clinical language models (CLMs) trained on clinical data have been used to improve performance in biomedical natural language processing tasks. In this work, we investigate the risks of training-data leakage through white-box or black-box access to CLMs. We design and employ membership inference attacks to estimate the empirical privacy leaks for model architectures like BERT and GPT2. We show that membership inference attacks on CLMs lead to non-trivial privacy leakages of up to 7%. Our results show that smaller models have lower empirical privacy leakages than larger ones, and masked LMs have lower leakages than auto-regressive LMs. We further show that differentially private CLMs can have improved model utility on clinical domain while ensuring low empirical privacy leakage. Lastly, we also study the effects of group-level membership inference and disease rarity on CLM privacy leakages.

[59] Zheng Li and Yang Zhang. "Membership Leakage in Label-only Exposures". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021, pp. 880–895. DOI: 10.1145/3460120.3484575. URL: https://dl.acm.org/doi/10.1145/3460120.3484575.
Machine learning (ML) has been widely adopted in various privacy-critical applications, e.g., face recognition and medical image analysis. However, recent research has shown that ML models are vulnerable to attacks against their training data. Membership inference is one major attack in this domain: Given a data sample and model, an adversary aims to determine whether the sample is part of the model's training set. Existing membership inference attacks leverage the confidence scores returned by the model as their inputs (score-based attacks). However, these attacks can be easily mitigated if the model only exposes the predicted label, i.e., the final model decision. In this paper, we propose decision-based membership inference attacks and demonstrate that label-only exposures are also vulnerable to membership leakage. In particular, we develop two types of decision-based attacks, namely transfer attack and boundary attack. Empirical evaluation shows that our decision-based attacks can achieve remarkable performance, and even outperform the previous score-based attacks in some cases. We further present new insights on the success of membership inference based on quantitative and qualitative analysis, i.e., member samples of a model are more distant to the model's decision boundary than non-member samples. Finally, we evaluate multiple defense mechanisms against our decision-based attacks and show that our two types of attacks can bypass most of these defenses.

[60] Kin Sum Liu, Bo Li, and Jie Gao. "Generative Model: Membership Attack, Generalization and Diversity". In: *CoRR, abs/1805.09898* (2018). DOI: 10.48550/arXiv.1805.09898. URL: https://arxiv.org/abs/1805.09898v1.
This paper considers membership attacks to deep generative models, which is to check whether a given instance x was used in the training data or not. Membership attack is an important topic closely related to the privacy issue of training data and most prior work were on supervised learning. In this paper we propose new methods to launch membership attacks against Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). The main idea is to train another neural network (called the attacker network) to search for the seed to reproduce the target data x. The difference of the generated data and x is used to conclude whether x is in the training data or not. We examine extensively the similarity/correlation and differences of membership attack with model generalization, overfitting, and diversity of the model. On different data sets we show our membership attacks are more effective than alternative methods.

[61] Kin Sum Liu et al. "Performing Co-membership Attacks Against Deep Generative Models". In: *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2019, pp. 459–467. DOI:

10.1109/ICDM.2019.00056. URL: https://ieeexplore.ieee.org/abstract/document/8970995.
In this paper we propose a new membership attack method called co-membership attacks against deep generative models including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). Specifically, membership attack aims to check whether a given instance x was used in the training data or not. A co-membership attack checks whether the given bundle of n instances were in the training, with the prior knowledge that the bundle was either entirely used in the training or none at all. Successful membership attacks can compromise the privacy of training data when the generative model is published. Our main idea is to cast membership inference of target data x as the optimization of another neural network (called the attacker network) to search for the latent encoding to reproduce x. The final reconstruction error is used directly to conclude whether x was in the training data or not. We conduct extensive experiments on a variety of datasets and generative models showing that: our attacker network outperforms prior membership attacks; co-membership attacks can be substantially more powerful than single attacks; and VAEs are more susceptible to membership attacks compared to GANs.

[62] Yiyong Liu et al. "Membership Inference Attacks by Exploiting Loss Trajectory". In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022, pp. 2085–2098. DOI: 10.1145/3548606.3560684. URL: https://yangzhangalmo.github.io/papers/CCS22-LossTrajectory.pdf.
Machine learning models are vulnerable to membership inference attacks in which an adversary aims to predict whether or not a particular sample was contained in the target model's training dataset. Existing attack methods have commonly exploited the output information (mostly, losses) solely from the given target model. As a result, in practical scenarios where both the member and non-member samples yield similarly small losses, these methods are naturally unable to differentiate between them. To address this limitation, in this paper, we propose a new attack method, called TrajectoryMIA, which can exploit the membership information from the whole training process of the target model for improving the attack performance. To mount the attack in the common black-box setting, we leverage knowledge distillation, and represent the membership information by the losses evaluated on a sequence of intermediate models at different distillation epochs, namely distilled loss trajectory, together with the loss from the given target model. Experimental results over different datasets and model architectures demonstrate the great advantage of our attack in terms of different metrics. For example, on CINIC-10, our attack achieves at least 6 times higher true-positive rate at a low false-positive rate of 0.1% than existing methods. Further analysis demonstrates the general effectiveness of our attack in more strict scenarios.

[63] Iyiola E Olatunji, Wolfgang Nejdl, and Megha Khosla. "Membership Inference Attack on Graph Neural Networks". In: *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE. 2021, pp. 11–20. DOI: 10.1109/TPSISA52974.2021.00002. URL: https://arxiv.org/abs/2101.06570.
Graph Neural Networks (GNNs), which generalize traditional deep neural networks on graph data, have achieved state-of-the-art performance on several graph analytical tasks. We focus on how trained GNN models could leak information about the member nodes that they were trained on. We introduce two realistic settings for performing a membership inference (MI) attack on GNNs. While choosing the simplest possible attack model that utilizes the posteriors of the trained model (black-box access), we thoroughly analyze the properties of GNNs and the datasets which dictate the differences in their robustness towards MI attack. While in traditional machine learning models, overfitting is considered the main cause of such leakage, we show that in GNNs the additional structural information is the major contributing factor. We support our findings by extensive experiments on four representative GNN models. To prevent MI attacks on GNN, we propose two

effective defenses that significantly decreases the attacker's inference by up to 60% without degradation to the target model's performance. Our code is available at https://github.com/iyempissy/rebMIGraph.

[64] Md Atiqur Rahman et al. "Membership Inference Attack against Differentially Private Deep Learning Model." In: *Trans. Data Priv.* 11.1 (2018), pp. 61–79. URL: http://www.tdp.cat/issues16/abs.a289a17.php.

The unprecedented success of deep learning is largely dependent on the availability of massive amount of training data. In many cases, these data are crowd-sourced and may contain sensitive and confidential information, therefore, pose privacy concerns. As a result, privacy-preserving deep learning has been gaining increasing focus nowadays. One of the promising approaches for privacy-preserving deep learning is to employ differential privacy during model training which aims to prevent the leakage of sensitive information about the training data via the trained model. While these models are considered to be immune to privacy attacks, with the advent of recent and sophisticated attack models, it is not clear how well these models trade-off utility for privacy. In this paper, we systematically study the impact of a sophisticated machine learning based privacy attack called the membership inference attack against a state-of-the-art differentially private deep model. More specifically, given a differentially private deep model with its associated utility, we investigate how much we can infer about the model's training data. Our experimental results show that differentially private deep models may keep their promise to provide privacy protection against strong adversaries by only offering poor model utility, while exhibit moderate vulnerability to the membership inference attack when they offer an acceptable utility. For evaluating our experiments, we use the CIFAR-10 and MNIST datasets and the corresponding classification tasks.

[65] Shahbaz Rezaei and Xin Liu. "On the Difficulty of Membership Inference Attacks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7892–7900. DOI: 10.1109/CVPR46437.2021.00780. URL: https://openaccess.thecvf.com/content/CVPR2021/papers/Rezaei_On_the_Difficulty_of_Membership_Inference_Attacks_CVPR_2021_paper.pdf.

Recent studies propose membership inference (MI) attacks on deep models, where the goal is to infer if a sample has been used in the training process. Despite their apparent success, these studies only report accuracy, precision, and recall of the positive class (member class). Hence, the performance of these attacks have not been clearly reported on negative class (non-member class). In this paper, we show that the way the MI attack performance has been reported is often misleading because they suffer from high false positive rate or false alarm rate (FAR) that has not been reported. FAR shows how often the attack model mislabel non-training samples (non-member) as training (member) ones. The high FAR makes MI attacks fundamentally impractical, which is particularly more significant for tasks such as membership inference where the majority of samples in reality belong to the negative (non-training) class. Moreover, we show that the current MI attack models can only identify the membership of misclassified samples with mediocre accuracy at best, which only constitute a very small portion of training samples. We analyze several new features that have not been comprehensively explored for membership inference before, including distance to the decision boundary and gradient norms, and conclude that deep models' responses are mostly similar among train and non-train samples. We conduct several experiments on image classification tasks, including MNIST, CIFAR-10, CIFAR-100, and ImageNet, using various model architecture, including LeNet, AlexNet, ResNet, etc. We show that the current state-of-the-art MI attacks cannot achieve high accuracy and low FAR at the same time, even when the attacker is given several advantages. The source code is available at https://github.com/shrezaei/MI-Attack.

[66] Ahmed Salem et al. "ML-leaks: Model and Data Independent Membership Inference Attacks

and Defenses on Machine Learning Models". In: *arXiv preprint arXiv:1806.01246* (2018). DOI: 10.48550/arXiv.1806.01246. URL: https://arxiv.org/abs/1806.01246.
Machine learning (ML) has become a core component of many real-world applications and training data is a key factor that drives current progress. This huge success has led Internet companies to deploy machine learning as a service (MLaaS). Recently, the first membership inference attack has shown that extraction of information on the training set is possible in such MLaaS settings, which has severe security and privacy implications. However, the early demonstrations of the feasibility of such attacks have many assumptions on the adversary, such as using multiple so-called shadow models, knowledge of the target model structure, and having a dataset from the same distribution as the target model's training data. We relax all these key assumptions, thereby showing that such attacks are very broadly applicable at low cost and thereby pose a more severe risk than previously thought. We present the most comprehensive study so far on this emerging and developing threat using eight diverse datasets which show the viability of the proposed attacks across domains. In addition, we propose the first effective defense mechanisms against such broader class of membership inference attacks that maintain a high level of utility of the ML model.

[67] Liwei Song, Reza Shokri, and Prateek Mittal. "Membership Inference Attacks Against Adversarially Robust Deep Learning Models". In: *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE. 2019, pp. 50–56. DOI: 10.1109/SPW.2019.00021. URL: https://ieeexplore.ieee.org/document/8844607.
In recent years, the research community has increasingly focused on understanding the security and privacy challenges posed by deep learning models. However, the security domain and the privacy domain have typically been considered separately. It is thus unclear whether the defense methods in one domain will have any unexpected impact on the other domain. In this paper, we take a step towards enhancing our understanding of deep learning models when the two domains are combined together. We do this by measuring the success of membership inference attacks against two state-of-the-art adversarial defense methods that mitigate evasion attacks: adversarial training and provable defense. On the one hand, membership inference attacks aim to infer an individual's participation in the target model's training dataset and are known to be correlated with target model's overfitting. On the other hand, adversarial defense methods aim to enhance the robustness of target models by ensuring that model predictions are unchanged for a small area around each sample in the training dataset. Intuitively, adversarial defenses may rely more on the training dataset and be more vulnerable to membership inference attacks. By performing empirical membership inference attacks on both adversarially robust models and corresponding undefended models, we find that the adversarial training method is indeed more susceptible to membership inference attacks, and the privacy leakage is directly correlated with model robustness. We also find that the provable defense approach does not lead to enhanced success of membership inference attacks. However, this is achieved by significantly sacrificing the accuracy of the model on benign data points, indicating that privacy, security, and prediction accuracy are not jointly achieved in these two approaches.

[68] Stacey Truex et al. "Towards Demystifying Membership Inference Attacks". In: *arXiv preprint arXiv:1807.09173* (2018). DOI: 10.48550/arXiv.1807.09173. URL: https://arxiv.org/abs/1807.09173.
Membership inference attacks seek to infer membership of individual training instances of a model to which an adversary has black-box access through a machine learning-as-a-service API. In providing an in-depth characterization of membership privacy risks against machine learning models, this paper presents a comprehensive study towards demystifying membership inference attacks from two complimentary perspectives. First, we provide a generalized formulation of the development of a black-box membership inference attack model. Second, we characterize the importance of model choice on model vulnerability

through a systematic evaluation of a variety of machine learning models and model combinations using multiple datasets. Through formal analysis and empirical evidence from extensive experimentation, we characterize under what conditions a model may be vulnerable to such black-box membership inference attacks. We show that membership inference vulnerability is data-driven and corresponding attack models are largely transferable. Though different model types display different vulnerabilities to membership inference, so do different datasets. Our empirical results additionally show that (1) using the type of target model under attack within the attack model may not increase attack effectiveness and (2) collaborative learning exposes vulnerabilities to membership inference risks when the adversary is a participant. We also discuss countermeasure and mitigation strategies.

[69] Stacey Truex et al. "Demystifying Membership Inference Attacks in Machine Learning as a Service". In: *IEEE Transactions on Services Computing* (2019). DOI: 10.1109/TSC.2019.2897554. URL: https://ieeexplore.ieee.org/document/8634878.
Membership inference attacks seek to infer membership of individual training instances of a model to which an adversary has black-box access through a machine learning-as-a-service API. In providing an in-depth characterization of membership privacy risks against machine learning models, this paper presents a comprehensive study towards demystifying membership inference attacks from two complimentary perspectives. First, we provide a generalized formulation of the development of a black-box membership inference attack model. Second, we characterize the importance of model choice on model vulnerability through a systematic evaluation of a variety of machine learning models and model combinations using multiple datasets. Through formal analysis and empirical evidence from extensive experimentation, we characterize under what conditions a model may be vulnerable to such black-box membership inference attacks. We show that membership inference vulnerability is data-driven and corresponding attack models are largely transferable. Though different model types display different vulnerabilities to membership inference, so do different datasets. Our empirical results additionally show that (1) using the type of target model under attack within the attack model may not increase attack effectiveness and (2) collaborative learning exposes vulnerabilities to membership inference risks when the adversary is a participant. We also discuss countermeasure and mitigation strategies.

[70] Lauren Watson et al. "On the Importance of Difficulty Calibration in Membership Inference Attacks". In: *arXiv preprint arXiv:2111.08440* (2021). DOI: 10.48550/arXiv.2111.08440. URL: https://arxiv.org/abs/2111.08440.
The vulnerability of machine learning models to membership inference attacks has received much attention in recent years. However, existing attacks mostly remain impractical due to having high false positive rates, where non-member samples are often erroneously predicted as members. This type of error makes the predicted membership signal unreliable, especially since most samples are non-members in real world applications. In this work, we argue that membership inference attacks can benefit drastically from *difficulty calibration*, where an attack's predicted membership score is adjusted to the difficulty of correctly classifying the target sample. We show that difficulty calibration can significantly reduce the false positive rate of a variety of existing attacks without a loss in accuracy.

[71] Ryan Webster et al. "This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces". In: *arXiv preprint arXiv:2107.06018* (2021). DOI: 10.48550/arXiv.2107.06018. URL: https://arxiv.org/abs/2107.06018.
Recently, generative adversarial networks (GANs) have achieved stunning realism, fooling even human observers. Indeed, the popular tongue-in-cheek website https://thispersondoesnotexist.com, taunts users with GAN generated images that seem too real to believe. On the other hand, GANs do leak information about their training data, as evidenced by membership attacks recently demonstrated in the literature. In this

work, we challenge the assumption that GAN faces really are novel creations, by constructing a successful membership attack of a new kind. Unlike previous works, our attack can accurately discern samples sharing the same identity as training samples without being the same samples. We demonstrate the interest of our attack across several popular face datasets and GAN training procedures. Notably, we show that even in the presence of significant dataset diversity, an over represented person can pose a privacy concern.

[72] Jiayuan Ye et al. "Enhanced Membership Inference Attacks Against Machine Learning Models". In: *arXiv preprint arXiv:2111.09679* (2021). DOI: 10.48550/arXiv.2111.09679. URL: https://arXiv.org/abs/2111.09679.

How much does a machine learning algorithm leak about its training data, and why? Membership inference attacks are used as an auditing tool to quantify this leakage. In this paper, we present a comprehensive *hypothesis testing framework* that enables us not only to formally express the prior work in a consistent way, but also to design new membership inference attacks that use reference models to achieve a significantly higher power (true positive rate) for any (false positive rate) error. More importantly, we explain *why* different attacks perform differently. We present a template for indistinguishability games, and provide an interpretation of attack success rate across different instances of the game. We discuss various uncertainties of attackers that arise from the formulation of the problem, and show how our approach tries to minimize the attack uncertainty to the one bit secret about the presence or absence of a data point in the training set. We perform a *differential analysis* between all types of attacks, explain the gap between them, and show what causes data points to be vulnerable to an attack (as the reasons vary due to different granularities of memorization, from overfitting to conditional memorization). Our auditing framework is openly accessible as part of the *Privacy Meter* software tool.